Preprints of the 21st IFAC World Congress (Virtual)
Berlin, Germany, July 12-17, 2020

An Optimal Linear Attack Strategy on Remote State Estimation

Hanxiao Liu [*,***] **Yuqing Ni** [**] **Lihua Xie** [*]
Karl Henrik Johansson [***]

[*] *School of Electrical and Electronic Engineering, Nanyang Technological University, Singapore (e-mail: hanxiao001@ntu.edu.sg, elhxie@ntu.edu.sg).*
[**] *Department of Electronic and Computer Engineering, Hong Kong University of Science and Technology, Clear Water Bay, Kowloon, Hong Kong (email: yniac@connect.ust.hk)*
[***] *School of Electrical Engineering and Computer Science, KTH Royal Institute of Technology, Sweden (email: kallej@kth.se)*

Abstract: This work considers the problem of designing an attack strategy on remote state estimation under the condition of strict stealthiness and ϵ-stealthiness of the attack. An attacker is assumed to be able to launch a linear attack to modify sensor data. A metric based on Kullback-Leibler divergence is adopted to quantify the stealthiness of the attack. We propose a generalized linear attack based on past attack signals and the latest innovation. We prove that the proposed approach can obtain an attack which can cause more estimation performance loss than linear attack strategies recently studied in the literature. The result thus provides a bound on the tradeoff between available information and attack performance, which is useful in the development of mitigation strategies. Finally, some numerical examples are given to evaluate the performance of the proposed strategy.

Keywords: Cyber-Physical Systems Security, State Estimation, Integrity Attacks.

1. INTRODUCTION

Cyber-Physical Systems (CPSs), which integrate computational elements and physical processes closely, are playing a more and more critical role in a large variety of fields which include transportation, power grid, military and environment. Most of them are of great importance to the normal operation of society and even to the whole nation. Any successful cyber-physical attacks will bring huge damages to critical infrastructure, human lives and properties, and even threaten the national security. Maroochy water breach in 2000 (Slay and Miller (2007)), Stuxnet malware in 2010 (Karnouskos (2011)), Ukraine power outage in 2015 (Whitehead et al. (2017)) and other security incidents, motivate us to pay more attention to the security of CPSs.

Recently, an enormous amount of research effort has been devoted to designing detection algorithms and secure state estimation strategies to enhance the security of CPSs. Mo and Sinopoli (2009) and Mo et al. (2015) analyzed the effect of replay attacks, where the attackers do not know the system information and replay the recorded measurements, and proposed a physical watermarking scheme to detect this kind of attacks. Liu et al. (2014) proposed the nuclear norm minimization approach and low rank matrix factorization approach to create a mechanism based on the properties of the nominal power grid to detect data injection attacks in a power grid. Teixeira et al. (2012) characterized the properties of zero dynamics attacks and provided necessary and sufficient conditions that the changes of inputs and outputs should satisfy to reveal attacks. Fawzi et al. (2014) proposed a novel characterization of the maximum number of attacks that can be detected and provided an algorithm motivated by compressed sensing to estimate the state with attacks.

To the best of our knowledge, the concept of stealthiness of the attack was first introduced as ϵ-stealthiness based on KL divergence in Bai et al. (2015, 2017b). The authors provided the corresponding ϵ-stealthy attack strategy to induce the maximal performance degradation for a scalar system through data injection. Kung et al. (2016) generalized the above results to vector systems and pointed out the differences between scalar systems and vector systems. Furthermore, Bai et al. (2017a) was devoted to seeking the optimal attack by compromising sensors' measurements. In this paper, we adopt the stealthiness metric employed in Bai et al. (2015, 2017a). Different from these works focusing on obtaining the maximal performance degradation and then deriving the corresponding attack strategy, we consider to maximize performance degradation given a specific linear attack type. Moreover, the performance degradation metric is slightly different from the above works.

⋆ This work is supported by the A*STAR Industrial Internet of Things Research Program, under the RIE2020 IAF-PP Grant A1788a0023, the Knut and Alice Wallenberg Foundation, the Swedish Foundation for Strategic Research, and the Swedish Research Council.

The linear integrity attack in our work was first proposed in Guo et al. (2016). An optimal linear attack policy was proposed to achieve the maximal performance degradation while not being detected. Some other extensions under different scenarios on this work could be found in Guo et al. (2019); Guo et al. (2017). Guo et al. (2018) also investigated this attack type in the detection framework based on KL divergence, which relaxed restrictions on false data detectors. However, since this type of attack only considers the latest information, it may not be optimal from the viewpoint of attacker. Motivated by this point, we consider a more general attack type which combines the past attack information and the latest innovation. Moreover, we focus on the sequence detection instead of one-slot detection.

This work considers the problem of designing a general linear attack strategy on remote state estimation under the condition of different stealthiness from the standpoint of the attacker. Our work builds on the above works and focuses on designing a more general linear attack strategy. The main contributions of this paper are threefold:

(1) We propose a more general linear attack type which employs the past attack data as well as the latest innovation and introduce the concept of ϵ-stealthy attacks to characterize the stealthiness level of an attack.
(2) We present the optimal attack strategy to achieve the maximal performance degradation for two specific attacks with different stealthiness.
(3) We prove that the proposed strategy performs better than the existing linear attack strategies in terms of performance degradation. Some numerical examples are provided to show this result.

Notations: $x_{k_1}^{k_2}$ is the sequence $\{x_{k_1}, x_{k_1+1}, \cdots, x_{k_2}\}$. The spectral radius $\rho(A) = \max\{|\lambda_1|, |\lambda_2|, \cdots, |\lambda_n|\}$, where $\lambda_1, \cdots, \lambda_n$ are the eigenvalues of the matrix $A \in \mathbb{R}^{n\times n}$. I_n denotes the identity matrix of order n.

2. PROBLEM FORMULATION

In this section, we introduce the system model as well as attack model. Besides, the stealthiness metric and performance degradation metric are provided to characterize the properties of attacks. Finally, we formulate the problem. The system diagram under consideration is illustrated in Fig. 1.

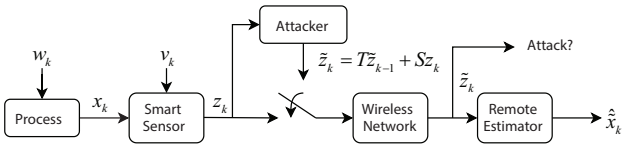

Fig. 1. The system diagram.

2.1 System Model

Let us consider a linear time-invariant (LTI) system described by the following equations:

$$x_{k+1} = Ax_k + w_k, \tag{1}$$
$$y_k = Cx_k + v_k, \tag{2}$$

where $x_k \in \mathbb{R}^n$ and $y_k \in \mathbb{R}^m$ are the state vector and all sensors' measurement at time k, respectively. $w_k \in \mathbb{R}^n$ denotes the process noise and $v_k \in \mathbb{R}^m$ is the measurement noise. $w_k \sim \mathcal{N}(0, Q)$ and $v_k \sim \mathcal{N}(0, R)$, where $Q \geq 0$ and $R > 0$, respectively. It is assumed that $w_0, w_1, \ldots$ and $v_0, v_1, \ldots$ are mutually independent.

Assumption 1. The spectral radius $\rho(A) < 1$ and the pair (A, C) is detectable and $(A, \sqrt{Q})$ is stabilizable.

The system is equipped with local smart sensors whose functions include signal conditioning, signal processing, and decision-making; see Lewis (2004). Here, we assume that the smart sensor employs the Kalman filter to process measurement and transmit the innovation to the remote estimator as follows:

$$\begin{aligned}
&\hat{x}_{k+1|k} = A\hat{x}_{k|k}, P_{k+1|k} = AP_{k|k}A^T + Q,\\
&K_k = P_{k|k-1}C^T(CP_{k|k-1}C^T + R)^{-1},\\
&\hat{x}_{k|k} = \hat{x}_{k|k-1} + K_k(y_k - C\hat{x}_{k|k-1}),\\
&P_{k|k} = P_{k|k-1} - K_kCP_{k|k-1},
\end{aligned}$$

with initialization $\hat{x}_{0|-1} = \bar{x}_0$.

It is known that the Kalman gain will converge exponentially due to Assumption 1. Hence, we consider a steady-state Kalman filter with gain K and the priori minimum mean square error (MMSE) P for the remaining of this paper where

$$P = \lim_{k\to\infty} P_{k|k-1}, \tag{3}$$
$$K = PC^T(CPC^T + R)^{-1}. \tag{4}$$

Hence, the Kalman filter can be rewritten as:

$$\hat{x}_{k+1|k} = A\hat{x}_{k|k}, \quad \hat{x}_{k|k} = \hat{x}_{k|k-1} + Kz_k,$$

where $z_k \triangleq y_k - C\hat{x}_{k|k-1}$ is the innovation of the Kalman filter at time k, which will be transmitted to the remote estimator and $z_k \sim \mathcal{N}(0, \sigma_z^2)$, where $\sigma_z^2 = CPC^T + R$.

Remark 2. In our problem formulation, we assume that the innovation is transmitted to the remote estimator via a wireless communication network. Note that $y_k = z_k + C\hat{x}_{k|k-1}$, which means z_k contains the same information as y_k. In the existing works such as Ribeiro et al. (2006), Guo et al. (2016), Li et al. (2017), and Guo et al. (2019), the sensor also sends innovation z_k to the remote estimator.

2.2 Attack Model

Next we introduce the attack model. The adversary is assumed to have the following capabilities:

(1) The attacker has access to all the real-time innovations from smart sensors.
(2) The attacker can modify the true innovation to arbitrary value in a specific form.
(3) The attacker has the knowledge of system matrix A.

Remark 3. The third capability could be relaxed. If the system parameter A is not known, the attacker can learn it by system identification.

The attacker records the real-time innovations from smart sensors and modifies them to $\tilde{z}_k$, i.e.,

$$\tilde{z}_k = T\tilde{z}_{k-1} + Sz_k, \tag{5}$$

where $T \in \mathbb{R}^{m\times m}$ and $S \in \mathbb{R}^{m\times m}$.

The remote estimator receives $\tilde{z}_k$ and updates the state estimate as follows:

$$\hat{\tilde{x}}_{k+1|k} = A\hat{\tilde{x}}_{k|k}, \quad \hat{\tilde{x}}_{k|k} = \hat{\tilde{x}}_{k|k-1} + K\tilde{z}_k.$$

Here, we initialize $\hat{\tilde{x}}_{0|-1} = \hat{x}_{0|-1}$ and $\tilde{z}_k = 0$ for $k \leq 0$.

2.3 Stealthiness Metric

From the perspective of attackers, they should be stealthy or do not want to be detected by the system detector, otherwise the system will design countermeasures against attacks. In this work, we employ a metric based on KL divergence measure to quantify the stealthiness of attack, which was first proposed in Bai et al. (2015).

Here, we propose the attack detection problem as a sequential hypothesis testing. The controller uses the received innovation sequence to carry out the following binary hypothesis testing:

$\mathcal{H}_0$: The remote estimator receives z_1^k.

$\mathcal{H}_1$: The remote estimator receives $\tilde{z}_1^k$.

In testing $\mathcal{H}_0$ versus $\mathcal{H}_1$, there are two types of errors that can be made: the first type is called "false alarm", which denotes that the estimator decides $\mathcal{H}_1$ given $\mathcal{H}_0$, and the second type is called "miss detection", which represents that the estimator decides $\mathcal{H}_0$ when $\mathcal{H}_1$ is correct. Here, we denote the probability of miss detection at time k as p_k^M, and the probability of false alarm is p_k^F. Furthermore, the probability of correct detection is p_k^D, which denotes that the controller decides $\mathcal{H}_1$ given $\mathcal{H}_1$. It is easy to know that $p_k^D + p_k^M = 1$. Two definitions about attack stealthiness level are provided as follows:

Definition 4. (Strictly stealthy attack (Bai et al., 2017a)). The attack is strictly stealthy if $p_k^F \geq p_k^D$ at time $k \geq 0$ holds for any detector.

Definition 5. (ϵ-stealthy attack (Bai et al., 2017a)). The attacker is ϵ-stealthy if

$$\limsup_{k\to\infty} -\frac{1}{k}\log p_k^F \leq \epsilon \tag{6}$$

holds for any detector that satisfies $0 < p_k^M < \delta$ for all times k, where $0 < \delta < 1$.

Remark 6. Definition 5 is motivated by Chernoff-Stein Lemma (see Cover and Thomas (2012)). This lemma shows that the best exponent in probability of error is given by the relative entropy. Please refer to Bai et al. (2017a) for more details.

2.4 Performance Degradation Metric

In this paper, we employ the ratio of the state estimation error covariance $\tilde{P}$ and P to quantify the performance degradation introduced by the attacker, i.e., $\eta = \dfrac{\mathrm{tr}\,\tilde{P}}{\mathrm{tr}\,P}$, where P is defined in (3) and $\tilde{P}$ is defined as follows:

$$\tilde{P} \triangleq \limsup_{k\to\infty} \frac{1}{k}\sum_{n=1}^{k} \tilde{P}_n, \tag{7}$$

where $\tilde{P}_n = E[(x_n - \hat{\tilde{x}}_{n|n-1})(x_n - \hat{\tilde{x}}_{n|n-1})^T]$ [1].

[1] Akin performance degradation metric could be found in Bai et al. (2015).

From the perspective of attackers, they need to design an appropriate attack strategy to maximize the ratio η, i.e.,

$$\arg_{T,S} \quad \limsup_{k\to\infty} \frac{\frac{1}{k}\sum_{n=1}^{k} \mathrm{tr}\,\tilde{P}_n}{\mathrm{tr}\,P}. \tag{8}$$

Remark 7. It is worth noticing that when there is no attack, $\tilde{z}_k = z_k$. As the initialization condition $\hat{\tilde{x}}_{0|-1} = \hat{x}_{0|-1}$, one can derive that $\hat{\tilde{x}}_{k|k-1} = \hat{x}_{k|k-1}$. Hence, $\tilde{P} = P$ and $\eta = 1$. In other words, the performance will not be degraded without attacks.

2.5 Problems of Interest

For the system described by (1) and (2) under attack type (5), we aim to tackle the following two optimization problems:

(1)

$$\begin{aligned} \max_{T,S} \quad & \limsup_{k\to\infty} \frac{\frac{1}{k}\sum_{n=1}^{k} \mathrm{tr}\tilde{P}_n}{\mathrm{tr}P}, \\ \text{s. t.} \quad & \text{The attack is strictly stealthy.} \end{aligned} \tag{9}$$

(2)

$$\begin{aligned} \max_{T,S} \quad & \limsup_{k\to\infty} \frac{\frac{1}{k}\sum_{n=1}^{k} \mathrm{tr}\tilde{P}_n}{\mathrm{tr}P}, \\ \text{s. t.} \quad & \text{The attack is } \epsilon\text{-stealthy.} \end{aligned} \tag{10}$$

We need to find the optimal attack pair (T^*, S^*) to induce the largest performance degradation while guaranteeing that the stealthiness level satisfies the corresponding requirement.

3. PRELIMINARY RESULTS

In order to quantify the stealthiness level of attacks, we need to employ the KL divergence (Kullback and Leibler (1951), Cover and Thomas (2012)), which is defined as:

Definition 8. (KL divergence). Let x_1^k and y_1^k be two random sequences with joint probability density functions $f_{x_1^k}$ and $f_{y_1^k}$, respectively. The KL divergence between x_1^k and y_1^k equals

$$D(x_1^k \| y_1^k) = \int_{-\infty}^{+\infty} \log \frac{f_{x_1^k}(\xi_1^k)}{f_{y_1^k}(\xi_1^k)} f_{x_1^k}(\xi_1^k) d\xi_1^k. \tag{11}$$

One can see that $D(x_1^k \| y_1^k) \geq 0$, and $D(x_1^k \| y_1^k) = 0$ if and only if $f_{x_1^k} = f_{y_1^k}$. Generally speaking, KL divergence is asymmetric, i.e., $D(x_1^k \| y_1^k) \neq D(y_1^k \| x_1^k)$.

The necessary and sufficient conditions for strictly stealthy attacks and ϵ-stealthy attacks are provided as follows [2]:

Lemma 9. (Condition for Strictly Stealthy attacks). (Bai et al. (2017a)) An attack sequence $\tilde{z}_1^\infty$ is strictly stealthy if and only if $\tilde{z}_1^\infty$ is a sequence of i.i.d. Gaussian random variables with zero mean and variance $Cov(z_k) = CPC^T + R$.

Lemma 10. (Conditions for ϵ-stealthy attacks). (Bai et al. (2017a)) If an attack $\tilde{z}_1^\infty$ is ϵ-stealthy, then

$$\limsup_{k\to\infty} \frac{1}{k} D(\tilde{z}_1^k \| z_1^k) \leq \epsilon.$$

[2] For more details about the proofs, please refer to Bai et al. (2017a).

Conversely, if an attack sequence $\tilde{z}_1^\infty$ is ergodic and satisfies $\lim_{k\to\infty}\frac{1}{k}D(\tilde{z}_1^k\|z_1^k)\leq\epsilon$, then the attack is ϵ-stealthy.

4. MAIN RESULTS

In this section, we will design an optimal attack strategy under strictly stealthy attacks and ϵ-stealthy attacks. For the sake of analysis, we focus on the scalar case, i.e., $m=n=1$. The vector case will be a potential future extension. The detailed solutions are provided in the following sections.

4.1 Strictly Stealthy Attack

The goal of this subsection is to design an optimal attack pair (T^*,S^*) of the optimization problem (9).

Theorem 11. For a strictly stealthy attack, the optimal attack pair of the optimization problem (9) is $(T^*,S^*)=(0,-1)$ and the corresponding performance degradation ratio is $\eta=1+\frac{4A^2K^2(C^2P+R)}{(1-A^2)P}$.

Proof. Firstly, according to the attack type (5) and the condition of strictly stealthy attacks in Lemma 9, it is easy to derive that the feasible solutions of the problem (9) are $(T,S)=(0,\pm1)$.

If $(T,S)=(0,1)$, $\tilde{z}_k=z_k$, which denotes that there is no attack in process, and the corresponding ratio $\eta=1$.

If $(T,S)=(0,-1)$, $\tilde{z}_k=-z_k$, which is aligned with the independence that strictly stealthy attack satisfies. Then we start to derive the corresponding ratio η. Rewrite $\tilde{P}_k$:

$$\begin{aligned}\tilde{P}_k=&E[(x_k-\hat{\tilde{x}}_{k|k-1})^2]\\=&P+E[(\hat{x}_{k|k-1}-\hat{\tilde{x}}_{k|k-1})^2]\\&+2E[(x_k-\hat{x}_{k|k-1})(\hat{x}_{k|k-1}-\hat{\tilde{x}}_{k|k-1})]\\\overset{(a)}{=}&P+E[(\hat{x}_{k|k-1}-\hat{\tilde{x}}_{k|k-1})^2].\end{aligned}\tag{12}$$

The reason why equation (a) holds can be found in Bai et al. (2017a). We do not elaborate it here.

Define $\tilde{e}_k\triangleq\hat{x}_{k|k-1}-\hat{\tilde{x}}_{k|k-1}$. One can derive that

$$\begin{aligned}E[\tilde{e}_k^2]=&E[(\hat{x}_{k|k-1}-\hat{\tilde{x}}_{k|k-1})^2],\\\overset{(b)}{=}&\sum_{n=1}^{k}A^{2n}K^2E\left[[z_{k-n}-(-z_{k-n})]^2\right]\\=&4A^2K^2(C^2P+R)\frac{1-A^{2k}}{1-A^2},\end{aligned}\tag{13}$$

where (b) holds because $\tilde{e}_0=\hat{x}_{0|-1}-\hat{\tilde{x}}_{0|-1}=0$.

Then, the covariance of the priori state estimate can be calculated as

$$\tilde{P}=\lim_{k\to\infty}\frac{1}{k}\sum_{n=1}^{k}\tilde{P}_n=P+\frac{4A^2K^2(C^2P+R)}{1-A^2},$$

and we obtain the performance degradation ratio:

$$\eta=\frac{P+\frac{4A^2K^2(C^2P+R)}{1-A^2}}{P}=1+\frac{4A^2K^2(C^2P+R)}{(1-A^2)P}>1.$$

Hence, the optimal attack pair of the optimization problem (9) is $(T,S)=(0,-1)$ and the corresponding performance degradation ratio is $\eta=1+\frac{4A^2K^2(C^2P+R)}{(1-A^2)P}$. ■

Remark 12. The above attack strategy is aligned with the results about the worst-case linear attack under the χ^2 false alarm detector obtained in Guo et al. (2016) and Bai et al. (2017a), which alter the sign of the innovation.

Remark 13. It is worth noticing that from (13), one can derive that $E[\tilde{e}_k^2]=0$ for any attack if $A=0$, and then $\eta=1$. Hence, we do not consider this case for the later discussion.

4.2 ϵ-stealthy Attack

The goal of this subsection is to design an optimal attack pair (T^*,S^*) to maximize the ratio η, i.e., to maximize $\tilde{P}$, under ϵ-stealthy attacks.

Lemma 14. For any T, the differential entropy of the sequence $\tilde{z}_1^k$ can be expressed as $\frac{k}{2}\log(2\pi eS^2\sigma_z^2)$.

Lemma 15. If the attack is ϵ-stealthy, then $|T|<1$.

From Lemma 15, for an ϵ-stealthy attack, we have

$$\lim_{k\to\infty}\frac{1}{k}D(\tilde{z}_1^k\|z_1^k)=-\frac{1}{2}-\frac{1}{2}\log(S^2)+\frac{S^2}{2(1-T^2)}\leq\epsilon,\tag{14}$$

where the third term of the above equation can be derived from the summation of geometric series.

Then, we consider the performance degradation for an ϵ-stealthy attack. An equivalent optimization problem is given in the following theorem, the proof of which can be found in the appendix for the sake of legibility.

Theorem 16. The optimization problem (10) is equivalent to the following problem:

$$\max_{T,S}\ J(T,S)=(1-S)^2+\frac{T^2S^2}{1-T^2}-\frac{2ATS(1-S-T^2)}{(1-T^2)(1-AT)},\tag{15a}$$

$$\text{s.t.}\ -\frac{1}{2}-\frac{1}{2}\log(S^2)+\frac{S^2}{2(1-T^2)}\leq\epsilon,\tag{15b}$$

$$|T|<1.\tag{15c}$$

Next we seek to obtain a solution, i.e., an optimal attack pair (T^*,S^*), of the above optimization problem. For the simplicity of notations, we use J to denote $J(T,S)$.

First, let us consider the constraint condition (15b). Fix $T=T_o$ and $\epsilon=\epsilon_o$ ($\epsilon_o\geq0$), S must satisfy:

$$-\frac{1}{2}-\frac{1}{2}\log(S^2)+\frac{S^2}{2(1-T_o^2)}\leq\epsilon_o.\tag{16}$$

Define

$$\mathscr{C}(S,T_0,\epsilon_0)\triangleq-\frac{1}{2}-\frac{1}{2}\log(S^2)+\frac{S^2}{2(1-T_0^2)}-\epsilon_0.\tag{17}$$

It can be derived that only when $T_o^2\leq1-e^{-2\epsilon_o}$, the inequality (16) has feasible solutions. And the solution lies in the interval $[-S_{\max},-S_{\min}]\bigcup[S_{\min},S_{\max}]$, where $S_{\max}$ and S_{min} are the largest and smallest positive solution to the equation $-\frac{1}{2}-\frac{1}{2}\log(S^2)+\frac{S^2}{2(1-T_o^2)}=\epsilon_o$, respectively.

Lemma 17. The optimal attack pair (T^*,S^*) must satisfy $-\frac{1}{2}-\frac{1}{2}\log(S^2)+\frac{S^2}{2(1-T^2)}=\epsilon$, where S^* is the corresponding smallest solution for $T=T^*$.

Remark 18. Consider the property of the objective function, the optimal solution for S is obtained when $S\leq0$.

Therefore, we only consider that $S \leq 0$ for the remaining of this work.

Lemma 19. When S is negative and the absolute value of T is fixed, $J(T,S) \geq J(-T,S)$, where the sign of T is the same as the sign of A.

Remark 20. For the simplicity of analysis, we only consider $T \geq 0$ and $A > 0$. Hence, T is non-negative in the above equation. The case when $T < 0$ and $A < 0$ is essentially the same.

Lemma 21. The optimization problem (15a) is equivalent to the following problem:

$$\max_{S,T} \; J(T,S) = (1-S)^2 + \frac{T^2S^2}{1-T^2} - \frac{2ATS(1-S-T^2)}{(1-T^2)(1-AT)}$$

$$\text{s.t.} \quad -\frac{1}{2} - \frac{1}{2}\log(S^2) + \frac{S^2}{2(1-T^2)} = \epsilon \tag{18a}$$

$$|T| \leq \sqrt{1-e^{-2\epsilon}}, \tag{18b}$$

Proof. It can be proved from (17) and Lemma 17. ■

Rewrite (18a) as follows:

$$T = f(S) \triangleq \sqrt{1 - \frac{S^2}{2\epsilon + 1 + \log(S^2)}}. \tag{19}$$

Insert (19) into (15a), one has:

$$J_1(S) = -(2\epsilon + \log(S^2)) - \frac{2S}{1-Af(S)} + \frac{2(2\epsilon+1+\log(S^2))}{1-Af(S)}$$
$$(-S_{o\max} \leq S \leq -e^{-\epsilon}),$$

where the range of S is derived from $0 \leq T^2 \leq 1 - e^{-2\epsilon}$.

Theorem 22. The solution to the above optimization problem is an optimal attack pair (T_{opt}, S_{opt}), where S_{opt} satisfies $\frac{\partial J_1}{\partial S}\big|_{S=S_{opt}} = 0$ and $T_{opt} = \sqrt{1 - \frac{{S_{opt}}^2}{2\epsilon+1+\log(S_{opt}^2)}}$. And the corresponding performance degradation ratio is $\eta = \frac{J_{opt}A^2K^2\sigma_z^2}{(1-A^2)P}$, where $J_{opt} = J(T_{opt}, S_{opt})$.

Proof. The main idea of the proof is to verify that the signs of the derivative of J with respect to S along the two boundaries are different, thus the optimal solution must exist in the feasible domain.

$$\frac{\partial J_1}{\partial S} = -2\frac{A^2f^2(S) + S - ASf(S) - 1}{S(1-Af(S))^2} - 2\frac{\left[S^2 - S(2\epsilon+1+\log S^2)\right]Af'(S)}{S(1-Af(S))^2}, \tag{20}$$

where $f'(S) = -\frac{\frac{S(2\epsilon+1+\log(S^2))-S}{(2\epsilon+1+\log(S^2))^2}}{\sqrt{1-\frac{S^2}{2\epsilon+1+\log(S^2)}}}$. One can prove that when $S \to -S_{omax}$, the derivative of J_1 is positive. And the derivative at $S = -e^{-\epsilon}$ is negative.

Since the function J_1 and the derivative of J_1 with respect to S are continuous, there must be at least one maximum point where its first derivative is zero. Hence, $\eta = \frac{J_{opt}A^2K^2\sigma_z^2}{(1-A^2)P}$, where $J_{opt} = J(T_{opt}, S_{opt})$. ■

Corollary 23. The proposed attack strategy induces a larger performance degradation than the existing linear attack strategy in Guo et al. (2018) under the same ϵ-stealthy attacks.

5. SIMULATION

In this section, we provide some numerical examples to evaluate the performance of the proposed attack strategy. We consider an LTI system and set $A = 0.4, C = 1, Q = 0.2$, and $R = 0.5$. It is easy to derive that $K = 0.3102$, and $P = 0.2248$. Here, we run 100 thousand simulations to average them. The ratio of the state estimation error

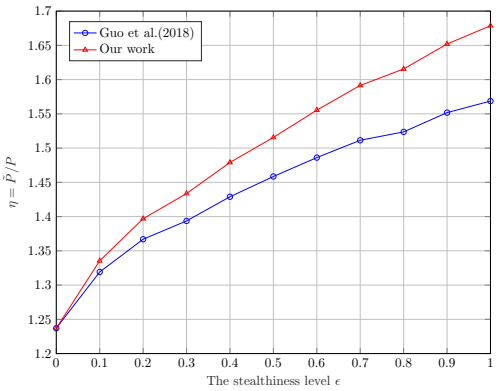

Fig. 2. The ratio η v.s. ϵ with fixed system parameters.

covariance $\tilde{P}$ to P v.s. stealthiness level ϵ is shown in Fig. 2. From this figure, one could see that the error covariance obtained in our work is equal to the one obtained in the exiting work Guo et al. (2018) when $\epsilon = 0$. And the error covariance obtained in our work is larger than the one derived in Guo et al. (2016). Furthermore, the difference of the error covariance between our work and Guo et al. (2018) gets larger as ϵ grows.

6. CONCLUSION

In this paper, an optimal linear attack strategy based on both the past attack signals and the latest innovation was proposed to achieve maximal performance degradation while guaranteeing a prescribed stealthiness level. For strictly stealthy attacks, the result derived in this paper is aligned with the existing work. For ϵ-stealthy attacks, we derived an optimal linear attack strategy and proved that the performance degradation of the optimal attack pair computed by using our proposed approach is better than the existing one. Simulation results were presented to support the theoretical results. For future works, we would like to generalize the results to a vector system, as well as analyze the performance of the optimal attack strategy under other performance metrics.

APPENDIX

Proof of Theorem 16: Rewrite $\tilde{e}_{k+1}$:

$$\tilde{e}_{k+1} = A\tilde{e}_k + AK(1-S)z_k - AKT\tilde{z}_{k-1}. \tag{21}$$

From (21), we have $E[\tilde{e}_k] = 0$ and $\tilde{e}_k$ is independent of z_k. Hence, the covariance of $\tilde{e}_k$ is:

$$\begin{aligned} &E[(\tilde{e}_{k+1})^2] \\ =&A^2E[(\tilde{e}_k)^2] + [AK(1-S)]^2\sigma_z^2 + (AKT)^2E[(\tilde{z}_{k-1})^2] \\ &+ 2A^2K(1-S)E[\tilde{e}_k z_k] - 2A^2KTE[\tilde{e}_k\tilde{z}_{k-1}] \\ \overset{(a)}{=}&A^2E[(\tilde{e}_k)^2] + [AK(1-S)]^2\sigma_z^2 + (AKT)^2E[(\tilde{z}_{k-1})^2] \\ &- 2A^2KTE[\tilde{e}_k\tilde{z}_{k-1}], \end{aligned} \tag{22}$$

where

$$\begin{aligned}\tilde{e}_k &= A\tilde{e}_{k-1} + AK(1-S)z_{k-1} - AKT\tilde{z}_{k-2} \\ &= A^k\tilde{e}_0 + AK\left[\sum_{i=0}^{k-1} A^i(1-S)z_{k-1-i} - \sum_{i=0}^{k-1} A^iT\tilde{z}_{k-2-i}\right],\end{aligned}$$

and Equation (a) holds due to the independence between $\tilde{e}_k$ and z_k. From (22), one can obtain that:

$$\lim_{k\to\infty}\frac{1}{k}E[(\tilde{e}_1)^2] = \lim_{k\to\infty}\frac{1}{k}\left[AK(1-S)\right]^2\sigma_z^2 \overset{(b)}{=} 0,$$

and

$$\begin{aligned}&\lim_{k\to\infty}\frac{1}{k}E[(\tilde{e}_{k+1})^2] \\ &\overset{(c)}{=} \lim_{k\to\infty}\frac{1}{k}\Big(A^2E[(\tilde{e}_k)^2] - 2A^2KTE[\tilde{e}_k\tilde{z}_{k-1}]\Big) \overset{(d)}{=} 0,\end{aligned}$$

where (b), (c) and (d) hold since $|A| < 1$, $|T| < 1$, K and σ_z^2 are constants, and S is bounded due to the property of (15a) and (15b).

Consider the asymptotic behavior for (22) and take the limit, we have:

$$\begin{aligned}&\lim_{k\to\infty}\frac{1-A^2}{k}\sum_{n=1}^{k}E[(\tilde{e}_{n+1})^2] \\ =&\left[AK(1-S)\right]^2\sigma_z^2 + \frac{(AKT)^2S^2}{1-T^2}\sigma_z^2 \\ &- 2A^3K^2TS\left[\frac{1-S}{1-AT} - \frac{T^2S}{(1-T^2)(1-AT)}\right]\sigma_z^2,\end{aligned}$$

Since $\sigma_z^2 > 0$, $A^2K^2 > 0$ and $P > 0$, one can simplify the above optimization as

$$\begin{aligned}\max_{S,T}\quad & (1-S)^2 + \frac{T^2S^2}{1-T^2} - \frac{2ATS(1-S-T^2)}{(1-T^2)(1-AT)}, \\ \text{s.t.}\quad & -\frac{1}{2} - \frac{1}{2}\log(S^2) + \frac{S^2}{2(1-T^2)} \le \epsilon, \\ & |T| < 1,\end{aligned}$$

where constraint conditions are from (14) and Lemma 15.

REFERENCES

Bai, C.Z., Gupta, V., and Pasqualetti, F. (2017a). On kalman filtering with compromised sensors: Attack stealthiness and performance bounds. *IEEE Transactions on Automatic Control*, 62(12), 6641–6648.

Bai, C.Z., Pasqualetti, F., and Gupta, V. (2015). Security in stochastic control systems: Fundamental limitations and performance bounds. In *2015 American Control Conference (ACC)*, 195–200. IEEE.

Bai, C.Z., Pasqualetti, F., and Gupta, V. (2017b). Data-injection attacks in stochastic control systems: Detectability and performance tradeoffs. *Automatica*, 82, 251–260.

Cover, T.M. and Thomas, J.A. (2012). *Elements of information theory*. John Wiley & Sons.

Fawzi, H., Tabuada, P., and Diggavi, S. (2014). Secure estimation and control for cyber-physical systems under adversarial attacks. *IEEE Transactions on Automatic control*, 59(6), 1454–1467.

Guo, Z., Shi, D., Johansson, K.H., and Shi, L. (2019). Worst-case innovation-based integrity attacks with side information on remote state estimation. *IEEE Transactions on Control of Network Systems*, 6(1), 48–59. doi: 10.1109/TCNS.2018.2793664.

Guo, Z., Shi, D., Johansson, K.H., and Shi, L. (2016). Optimal linear cyber-attack on remote state estimation. *IEEE Transactions on Control of Network Systems*, 4(1), 4–13.

Guo, Z., Shi, D., Johansson, K.H., and Shi, L. (2017). Consequence analysis of innovation-based integrity attacks with side information on remote state estimation. *IFAC-PapersOnLine*, 50(1), 8399–8404.

Guo, Z., Shi, D., Johansson, K.H., and Shi, L. (2018). Worst-case stealthy innovation-based linear attack on remote state estimation. *Automatica*, 89, 117–124.

Karnouskos, S. (2011). Stuxnet worm impact on industrial cyber-physical system security. In *IECON 2011-37th Annual Conference of the IEEE Industrial Electronics Society*, 4490–4494. IEEE.

Kullback, S. and Leibler, R.A. (1951). On information and sufficiency. *The annals of mathematical statistics*, 22(1), 79–86.

Kung, E., Dey, S., and Shi, L. (2016). The performance and limitations of *epsilon*-stealthy attacks on higher order systems. *IEEE Transactions on Automatic Control*, 62(2), 941–947.

Lewis, F.L. (2004). Wireless sensor networks. *Smart environments: technologies, protocols, and applications*, 11–46.

Li, Y., Shi, L., and Chen, T. (2017). Detection against linear deception attacks on multi-sensor remote state estimation. *IEEE Transactions on Control of Network Systems*, 5(3), 846–856.

Liu, L., Esmalifalak, M., Ding, Q., Emesih, V.A., and Han, Z. (2014). Detecting false data injection attacks on power grid by sparse optimization. *IEEE Transactions on Smart Grid*, 5(2), 612–621.

Mo, Y. and Sinopoli, B. (2009). Secure control against replay attacks. In *2009 47th Annual Allerton Conference on Communication, Control, and Computing (Allerton)*, 911–918. IEEE.

Mo, Y., Weerakkody, S., and Sinopoli, B. (2015). Physical authentication of control systems: Designing watermarked control inputs to detect counterfeit sensor outputs. *IEEE Control Systems Magazine*, 35(1), 93–109.

Ribeiro, A., Giannakis, G.B., and Roumeliotis, S.I. (2006). SOI-KF: Distributed Kalman filtering with low-cost communications using the sign of innovations. *IEEE Transactions on signal processing*, 54(12), 4782–4795.

Slay, J. and Miller, M. (2007). Lessons learned from the maroochy water breach. In *International Conference on Critical Infrastructure Protection*, 73–82. Springer.

Teixeira, A., Shames, I., Sandberg, H., and Johansson, K.H. (2012). Revealing stealthy attacks in control systems. In *2012 50th Annual Allerton Conference on Communication, Control, and Computing (Allerton)*, 1806–1813. IEEE.

Whitehead, D.E., Owens, K., Gammel, D., and Smith, J. (2017). Ukraine cyber-induced power outage: Analysis and practical mitigation strategies. In *2017 70th Annual Conference for Protective Relay Engineers (CPRE)*, 1–8. IEEE.